Statystyka II Mat. Egzamin komputerowy 13.06.08

1. Porównaj metody klasyfikacji *lda*, *qda*, *glm* (regresja logistyczna) w eksperymencie kroswalidacji pięciokrotnej na danych `MathAchSchool[,-c(1,6)]` z biblioteki nlme. Cecha przewidywana `y` przyjmuje dwie wartości: „Catholic" i „Public". Jako miarę efektywności predykcji przyjmij współczynnik Goodmana-Kruskala między `y` oraz `y_pred`.

2. Wykorzystując rozkład X=QR, bez użycia pętli, bez mnożenia i odwracania macierzy, napisz własną wersję funkcji `hatvalues`. Argumentem funkcji ma być tylko macierz cech obserwowalnych (predyktorów) X.

3. Wczytaj z archiwum *MacierzePot3.zip* 12 potencjałów kontaktowych (PK). PK są to symetryczne macierze 20 x 20 opisujące oddziaływania 210 par aminokwasów w białkach. Napisz ogólną funkcję lower`Vec,` która przekształca macierz na wektor złożony z elementów należących do jej dolnej części bez przekątnej. Ze zbioru wektorów $\boldsymbol{x}_1$, …, $\boldsymbol{x}_{12}$ otrzymanych z 12 PK, zbuduj macierz danych X=[$\boldsymbol{x}_1$, …, $\boldsymbol{x}_{12}$] (190 obserwacji 12-wymiarowych). Za pomocą funkcji `boxplot.stat` sprawdź, czy w zbiorze X są obserwacje odstające?

4. Wykonaj klasteryzację hierarchiczną `hclust` metodą "`median`" danych z poprzedniego zadania. Dendrogram zapisz do pliku .e*ps*. Na podstawie wykresu separowalności i sylwetki oszacuj liczbę klastów. Napisz ogólną funkcję `reduceDim(X,proc)`, która przekształca macierz danych X na macierz złożoną z jej k-pierwszych składowych głównych, gdzie k jest minimalną liczbą składowych, których łączna wariancja stanowi przynajmniej `proc` procent całkowitej wariancji X. Jakie jest k dla `proc`=80 ?

5. Zbuduj model regresji liniowej wielu zmiennych dla długości życia `Life.Exp` wykorzystując pozostałe cechy ze zbioru danych `state.x77`:
`X=data.frame(state.x77,row.names=state.abb,check.names=T)`